Supporting Credibility Assessment of News in Social Media using Star Ratings and Alternate Sources

Michael P. Kenning
Dept. of Computer Science
Swansea University
Swansea, SA2 8PP, UK
michael.p.kenning@bath.edu

Ryan Kelly
Computing and Information Systems
The University of Melbourne
Melbourne, VIC, 3060, Australia
ryan.kelly@unimelb.edu.au

Simon L. Jones
Dept. of Computer Science
University of Bath
Bath, BA2 7AY, UK
s.l.jones@bath.ac.uk

Abstract
This paper reports findings from a preliminary experiment in which we designed and tested two interface augmentations for enhancing credibility judgments of news stories on Facebook. We find that users' credibility judgments can be improved by the two augmentations, though the changes in credibility scores were not statistically significant. However, participants spent longer using the design that gave them control over the evaluation process, and appeared to be more confident about the choices they made using it—despite the fact that their judgments were actually less accurate. We outline directions for future work based on these findings.

Author Keywords
Credibility Assessment, News Articles, Social Media

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Introduction
Recent years have seen the Internet emerge as a leading platform for the consumption of news content. Studies suggest that people are increasingly reading news online, with one reporting that two-thirds of US adults now acquire news via social media platforms such as Twitter

Figure 1:
Design A: The standard representation of a news item on Facebook, as was used in our experiment.

Figure 2:
Design B: A Facebook news item shown with the *Star Ratings* augmentation.

and Facebook [7]. However, people also report a lack of trust in these sites as sources for obtaining news [6]. A key concern here is the perceived *credibility* of the information they provide [8], an issue that has recently been brought to the fore by an apparent proliferation of biased, misleading or outright false news [2]. This raises the question of how users can be supported in judging the credibility of news articles they encounter on social media.

The HCI literature contains a number of solutions for supporting users' credibility judgments. A common feature of this work is that it aims to support users by augmenting web content with additional information that was previously latent or unavailable. For example, Schwarz and Morris [8] found that the perceived credibility of search engine returns could be enhanced by displaying meta-information about the kinds of people visiting a page (e.g. visits by experts vs. non-experts). Wang et al. [9] explored the potential for search engines to verify facts by presenting users with search results about questionable claims made on websites. Other work has explored the potential for users to provide subjective credibility ratings of webpages to support the judgments of future readers [4] and for such ratings to be obtained through automated techniques [1]. Similarly, services such as TweetCred[1] offer algorithmically generated assessments in the form of a numeric 'credibility score' [3] that can provide users with an additional resource for judging the credibility of information found online.

However, none of this previous work has examined techniques for supporting credibility judgments of news articles on social media. The present research begins to address this gap by studying the efficacy of two lightweight user interface augmentations (see Figures 2 and 3), using news on Facebook as an exemplary case. Our aim is not to improve the design of Facebook, but is rather to use it as a vehicle for exploring the larger question of how to enhance users' confidence in the information they acquire via social media. Our study contributes early evidence to show that the augmentations might be useful for supporting users' credibility assessments, and draws attention to challenges related to user control over the evaluation process, effort, and satisfaction with the user experience. It also opens up directions for future work in this area.

[1] http://twitdigest.iiitd.edu.in/TweetCred/

User Study: Method

We created an experiment to compare two approaches for supporting credibility judgments of news stories in social media. The first approach, *Star Rating* (Figure 2), augments a news item with a rating scale that mimics the way in which credibility has been supported elsewhere in the literature; that is, through a minimal cue which is apparently reliable but which calls on the user to make a decision about its utility [3]. The second approach, *Alternate Sources* (see Figures 3 and 4), augments a news item with a button that reveals a new pane next to the story. The window contains a list of alternate sources which also report on the subject of the news item. By providing a list of alternate sources, we envisage that users will be better placed to evaluate the credibility of the main article by comparing it to other articles on the same subject. Furthermore, we see both of these designs as feasible for deployment within Facebook, which we hoped would support realistic engagement with them by participants in our experiment.

The above designs were compared to a basic representation of a Facebook news item, presented to participants in the form of a mockup designed to mirror

Figure 3:
Design C: A Facebook news item shown with the *Alternate Sources* augmentation.

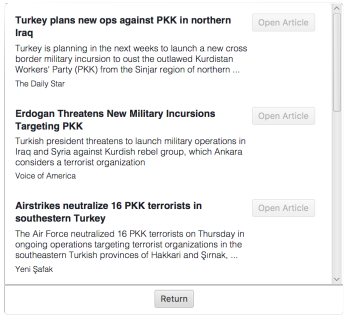

Figure 4:
Pressing the "Who else has reported on this story?" button (see Figure 3, above) reveals a pane containing a list of alternate sources that also report on the story.

the way in which Facebook displayed news articles at the time of our study (see Figure 1). The experiment used a within-subjects design in which participants used the designs to judge the credibility of seven real news articles. Our interest was in whether the designs might bring about changes in users' credibility ratings, and which of the two designs might be preferred by potential users.

To obtain a set of articles to test the designs, we surveyed news items that were trending on Google News between the 5th and 8th of April 2017 (shortly before our experiment took place). In our selection we avoided articles that covered:
1) Controversial topics that may generate highly-polarised opinions in our study site (the United Kingdom), e.g. the election of Donald Trump or Brexit.
2) Esoteric matters, e.g. recent advances in Particle Physics, which may make participants feel that they know too little to make a judgment or overconfident if they happen to know a lot about such an obscure topic.
3) Topics with irrefutable claims or those based on commonly held knowledge. For example, 'Christmas Day is the 25th of December' is a headline that most people would know to be true and therefore not require any evaluation of its credibility.

Our searches yielded seven suitable news items that covered a variety of current affairs from around the globe (see Table 1). These items were selected based on the considerations outlined above, alongside the availability of multiple sources reporting on the topic (at least 7 for each). We selected one article from the results on each topic to form a set of 'main articles', i.e. the target news items presented to participants in the experiment.

To support the implementation of the *Star Ratings* design, credibility scores were obtained from TweetCred by submitting the URLs for each of the main articles as a search on Twitter. The modal rating at the time of the search was taken as the star rating used in the experiment. For the *Alternate Sources* design, the participants were provided with between 7 and 9 (mode = 8) alternative articles for each news item. These articles were acquired using the same search terms used to find the main article.

Article Presentation and Ordering
The experiment was conducted using a custom Java application, with all of the designs implemented as mockups, i.e. they mirrored the aesthetic appearance of Facebook but were not actually deployed to participants' Facebook News Feeds. All of the study materials were presented to participants using a laptop computer.

To control for order effects, we based our experimental design on Schwarz and Morris' [8] earlier study of credibility assessment on websites. Participants in our experiment first saw the seven target stories through the standard Facebook design (Figure 1). The articles were shown individually (one per screen) and appeared in a randomized order. No other Facebook content was displayed. The same articles were then presented separately for each of the modified designs. No two participants saw the articles in the same order.

To account for order effects associated with the designs, half of the participants saw the designs in the order A-B-C (Standard; Star Ratings; Alternate Sources) and the other half saw them in the order A-C-B. We did not employ full counterbalancing because it makes little sense for participants to see a basic Facebook representation *after* one of the augmented designs. (The augmented designs would have ideally given participants richer insight into the credibility of a given article.) This also follows the procedure established by Schwarz & Morris [8].

#	Headline	Source
1	Turkey plans military incursion into Iraq: Report	Press TV
2	Russia interfering in French, German elections	Politico
3	David Cameron's government accused of lobbying for Uber	The Sun
4	Greek bailout talks could fail within DAYS as Tsipras threatens to run to EU members	Daily Express
5	Scotland May Do Better on Its Own as EEA Member, Analyst Says	Sputnik
6	Turbulence could increase dramatically thanks to more CO2 in the air	Popular Mechanics
7	Unilever may lose one headquarters and be British not Dutch	Dutch-News.nl

Table 1: List of the seven main articles used in the experiment. Between 7–9 alternate sources were obtained for each of these articles.

Procedure

Convenience and snowball sampling were used to recruit participants using adverts posted on Facebook. Ten people (6 male, 4 female) volunteered for the study. (Table 2 provides more information about the sample.) The experiment was conducted in a quiet laboratory at the University of Bath. All procedures were designed in accordance with our institution's code of ethics.

Participants arrived at the study individually, signed a consent form and completed a demographic questionnaire. The experimental software was pre-loaded on a laptop computer. Each participant first completed a training phase to clarify questions about the designs. The experimenter then left the room and the participant was taken through the seven articles by the software, presented using the standard Facebook design.

For each article, participants provided three ratings: perceived credibility, perceived bias, and an assessment of confidence in these ratings. All ratings were 1–5, where 1 = Not at all [credible/biased/confident] and 5 = Absolutely [credible/biased/confident]. After finishing with the standard design, the participant moved on to either Design B (Star Rating) or C (Alternate Sources), depending on the condition to which they had been assigned. After assessing the articles in this design, the participant was taken to the final design. All user input was logged for later analysis. We also recorded the time spent assessing each article to explore whether participants spent longer using a particular design.

At the end of the experiment, the first author conducted a short semi-structured interview (maximum 5 minutes) to capture participants' thoughts about the three designs. Interviews were audio recorded and transcribed after the study. Participants were debriefed about the research.

Results

Credibility Judgments

To examine the impact of the designs on credibility judgments, we first measured the distance (the absolute difference) between users' credibility ratings and a ground truth measure for each of the main articles.[2]

The results indicated that the two augmentations did lead to slight improvements to participants' credibility judgments. The mean distance between credibility and ground truth in the Standard news representation was 1.11; in the Alternate Sources design this distance improved to 1, and in the Star Ratings design it was 0.76. However, a Friedman test showed that changes were not statistically significant, $\chi^2(2) = 3.84$, $p = 0.147$. We also explored the frequency of exact agreement between participants' ratings and the ground truth score. The average agreement was 0.33 for the Standard design; 0.25 in Alternate Sources; and 0.47 in Star Rating. These differences were not statistically significant, Friedman $\chi^2(2) = 5.36$, $p = 0.169$. A Friedman test also showed that the designs had no significant impact on participants' ratings of article bias, $\chi^2(2) = 1.316$, $p = 0.518$.

Confidence in Judgments

A Friedman test revealed a statistically significant difference between participants' confidence ratings in each design, $\chi^2(2) = 7.947$, $p = 0.019$. Post hoc analysis conducted using the Wilcoxon signed-rank test revealed a statistically significant difference between the confidence

[2]Here, we took the modal credibility score provided by TweetCred as an indicator of ground truth for each article. The downside of this approach is that participants could have achieved ground truth by simply copying what was shown by the Star Rating. However, we did not tell them to do this, and the fact participants' scores did not align perfectly with ground truth in this condition indicates that participants did not do this in practice.

Participant Information
Demographics The age of participants ranged from 18–23 years (Mean = 21.4). All participants were British.
Facebook Usage All participants had Facebook accounts. Two participants reported to have used Facebook for 5 years or less; the rest reported to have used it for 10 years or less. Seven participants stated that they use Facebook every day; two stated that they use it every few days; the last stated that they use it at least once per week.
Facebook for News Four participants claimed to get a 'considerable amount' of their news via Facebook. The remaining six stated that they receive 'very little' of their news via Facebook.

Table 2: Information about participants' demographics, Facebook use, and use of Facebook to acquire news.

ratings for Standard (Mean = 4.00) and Alternate Sources (Mean = 4.386), Z = -2.108, p = 0.037, and between Star Ratings (Mean = 3.986) and Alternate Sources, Z = -2.608, p = 0.007. There was no significant difference in confidence ratings between the Standard design and Star Ratings, Z = -0.103, p = 0.918.

Time Spent Evaluating Articles
Paired-samples t-tests showed significant differences between the three designs for the time spent evaluating articles. Participants spent the most time evaluating articles when using Alternate Sources (Mean = 247 seconds). This was significantly different to their time spent using the Standard design (Mean = 192 seconds), t(9) = 2.599, p = 0.029, and to their time spent using Star Ratings (Mean = 109 seconds), t(9) = 7.099, p < 0.01. The difference between Standard and Star Ratings was also significantly different, t(9) = 6.417, p < 0.01.

Discussion

Our preliminary study provides an initial indication that participants' credibility judgments might be usefully shifted by the two designs we proposed. Compared to a standard Facebook representation, we observed slight improvements in ratings of article credibility, as indicated by the average distance from ground truth, when participants used the Star Ratings and Alternate Sources designs. However, these differences were not significantly different, suggesting that more research is required to determine the effectiveness of these two interventions.

Despite this limitation, our study provides several directions for future work. During our analysis, we found that the Star Ratings design produced slightly more accurate credibility judgments on average. This may initially seem unsurprising given that the design conveyed the measure of ground truth used in our analysis. However, the fact that participants appear willing to rely on the Star Rating for their credibility judgment is interesting in light of their post-experiment comments about the design. In particular, some participants expressed mistrust in the scoring system: *"It doesn't really tell you a lot"* (P4), *"I don't trust the stars. Who gave the stars and why?"* (P2). One participant suggested the star rating would be more *"relatable if there was [sic] several metrics on the star rating across several categories"* (Participant 3). In other words, participants expressed skepticism about this augmentation and yet still appeared to be willing to rely on it for a credibility judgment.

In contrast, remarks about Alternate Sources were more positive. One participant saw it as *"the best by far in terms of giving context"* (Participant 6) and that, rather than having to *"go off the language of the headline and the imagery"*, Alternate Sources permitted greater latitude in *"making your own evaluation"* (both quotes P7). These remarks dovetail with the finding that participants spent longer using this design and were more confident in their final judgments using it, which is intriguing given that their eventual credibility judgments were further from ground truth. Thus, participants may have felt as though they made a more thorough evaluation when using Alternate Sources (and felt more satisfied with their experience) but their eventual credibility assessment was less accurate than when they relied on the simple heuristic provided by the Star Rating. While this result hinges on the reliability of the ground truth measure used in our analysis (i.e. the modal credibility score provided by TweetCred), it should stimulate further work on the most appropriate means of supporting credibility judgments of news in social media. Our participants rationalised Alternate Sources as better because it provided an

opportunity to evaluate sources for themselves. Future work in this area should explore how objective measures of article credibility can be reconciled with user-driven assessments of news on social media.

Participants also offered comments into the perceived usability of the designs, and although Alternate Sources was valued, some participants considered the design to be *"more difficult to interpret since you have to flick through the other stories"* (P1). In contrast, participants described the Star Rating as *"very simple to interpret"* (P1), *"provid[ing] a hint as to how credible or biased the information is"* (P9). This suggests that there may be a trade-off between effort and usability in credibility evaluation, warranting study of this issue in future work.

Limitations & Future Work

Our study is preliminary and the number of participants is currently insufficient to draw reliable conclusions about how to support users' credibility judgments. (The low number of participants likely contributed to the lack of statistical significance in our findings.) Credibility is also known to be impacted by features of the source, e.g. author or publication venue [5], and our study did not explore this issue. Another factor that could be explored is the level of agreement between a target article and the alternate sources presented to the user. It seems logical to assume that users' perceptions might be changed if they see a list of alternate sources that disagree with the target article. Research could explore how users might be biased by such features, both for good and ill, in order to better inform future design activity in this space.

References

[1] Aggarwal, S., Van Oostendorp, H., Reddy, Y. R., and Indurkhya, B. Providing Web Credibility Assessment Support. In *Proceedings of the 2014 European Conference on Cognitive Ergonomics*, ACM (2014).

[2] Allcott, H., and Gentzkow, M. Social media and fake news in the 2016 election. *Journal of Economic Perspectives 31*, 2 (2017), 211–236.

[3] Gupta, A., Kumaraguru, P., Castillo, C., and Meier, P. Tweetcred: A real-time web-based system for assessing credibility of content on twitter. *CoRR abs/1405.5490* (2014).

[4] Huang, Z., Olteanu, A., and Aberer, K. CredibleWeb: a platform for web credibility evaluation. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, ACM (2013), 1887–1892.

[5] Morris, M. R., Counts, S., Roseway, A., Hoff, A., and Schwarz, J. Tweeting is believing?: Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, ACM (New York, NY, USA, 2012), 441–450.

[6] Pew Research Center. The modern news consumer, 2016. Accessed 11th January 2018. http://www.journalism.org/2016/07/07/the-modern-news-consumer/.

[7] Pew Research Center. News use across social media platforms, 2017. Accessed 11th January 2018. http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/.

[8] Schwarz, J., and Morris, M. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2011), 1245–1254.

[9] Wang, T., Zhu, Q., and Wang, S. Multi-verifier: A novel method for fact statement verification. *World Wide Web 18*, 5 (2015), 1463–1480.